Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Datenmanagement & -analyse

Überblick & Einführung

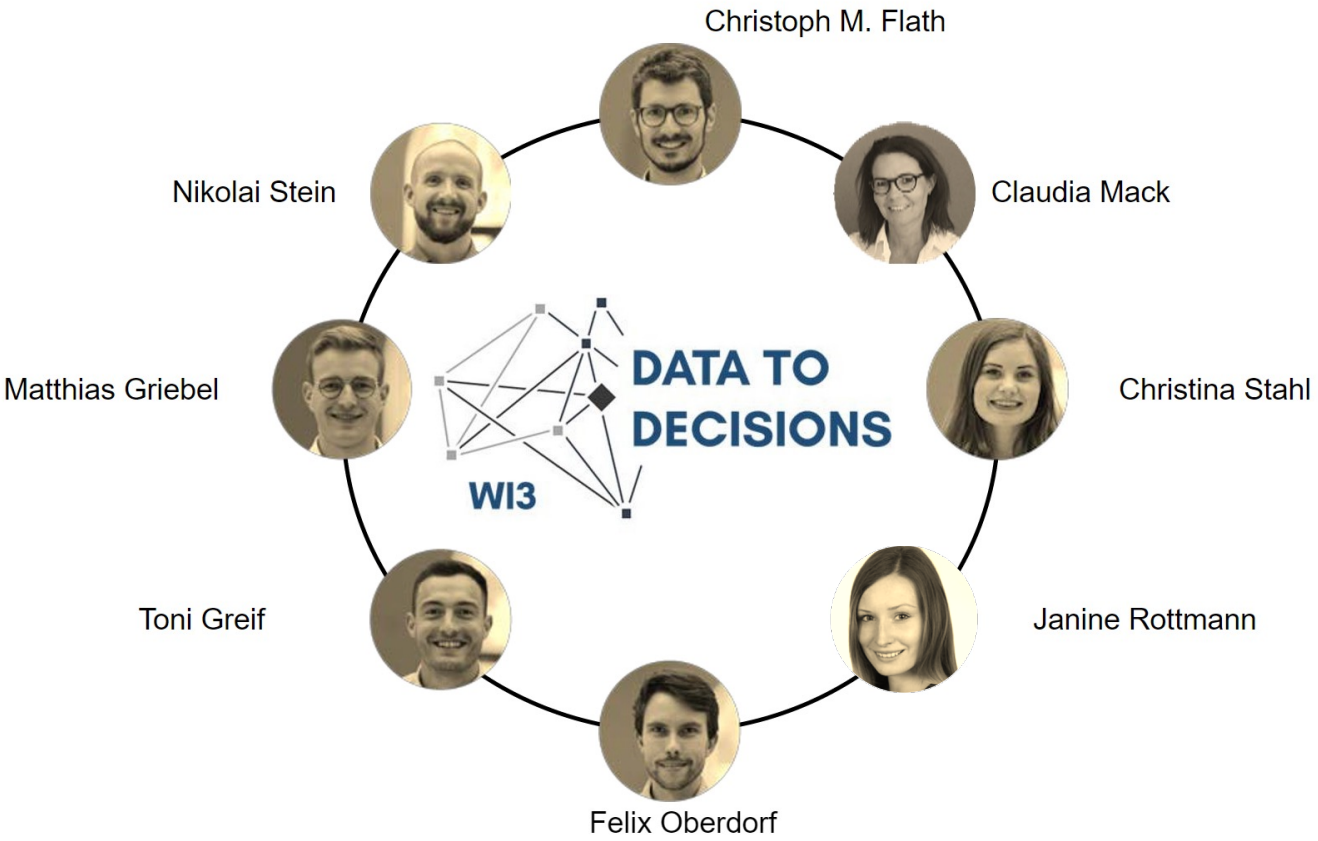
Prof. Dr. Christoph M. Flath
Lehrstuhl für WI & BA
Julius-Maximilians-Universität Würzburg

Sommersemester 2021

Agenda

1. **Organisatorisches**
2. **Motivation**
3. **Themenüberblick**

BWL12 – WI & Business Analytics



- Forschung und Lehre zu Methoden und Anwendungen von Business Analytics, insbesondere mathematische Optimierung und maschinelles Lernen
- Schwerpunktanwendungs-bereiche sind Industrie 4.0, Energiewirtschaft und Mobilität

Über uns

Christoph M. Flath

- Seit 04/2018 Lehrstuhlinhaber WI & Wirtschaftsinformatik
  - Forschungsthemen: Analytics, Industrie 4.0, Smart Grid, Future Mobility
- 06/2014-03/2018 Juniorprofessur für Operations Management (JMU)
- 05/2009-06/2014 Promotion und PostDoc (KIT)
- 10/2003-11/2008 Diplomstudiengang Wirtschafts-ingenieurwesen (KIT)

Dr. Nikolai Stein

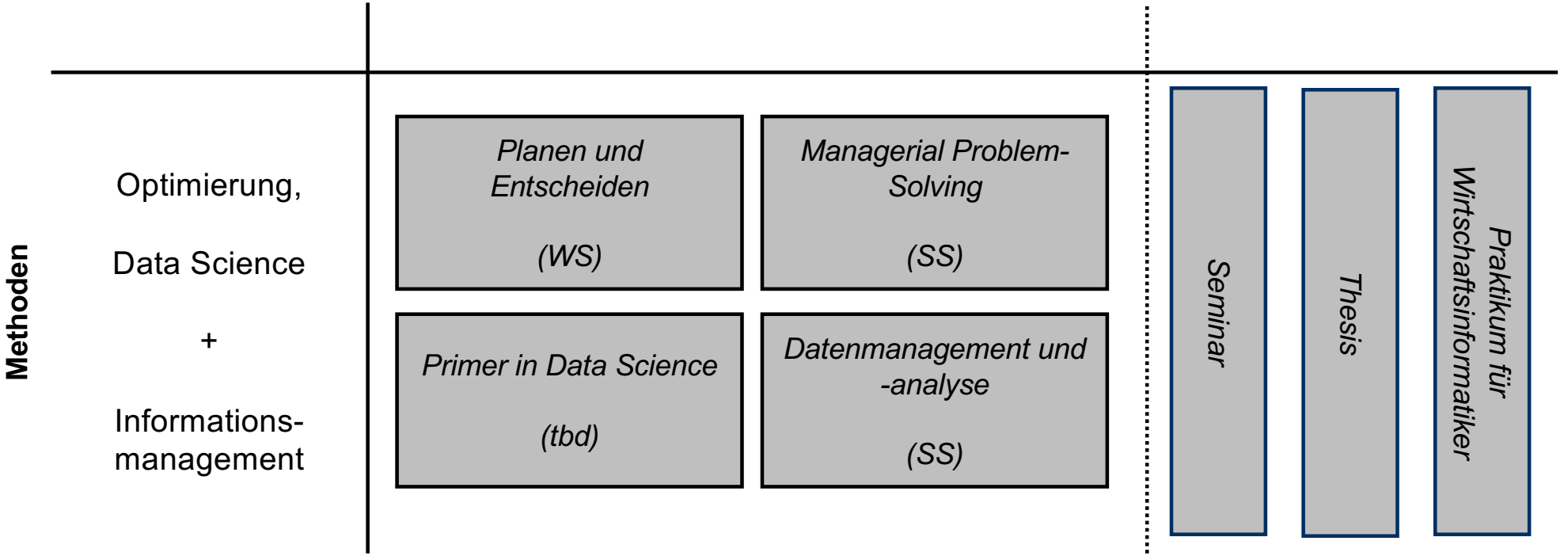
- Seit 01/2020 PostDoc am Lehrstuhl
- 05/2015 – 12/2019 Promotion @ am Lehrstuhl WI3
  - Research topics: Advanced Analytics and Industry 4.0
- Ausbildung
  - Bachelor (Wiwi, JMU)
  - Master (WI, JMU)

Up to you – bitte die Umfrage ausfüllen

Organisation der Veranstaltung

- Vorlesung Mittwoch, Übung Donnerstag
- Die Übung zielt auf die 1:1 Anwendung des gelernten ab
  - Kennenlernen von Abfragesprachen (dplyR, SQL)
  - Umgang mit Docker-Umgebungen
  - Selbständiges Umsetzen von Datenanalyseprojekten
- Veranstaltung ersetzt das alte Modul „Datenmodellierung“
  - Erweiterung um Ideen der Datenanalyse
- Ein paar Warnungen
  - Es geht nicht ohne Code
  - Veranstaltung findet zum ersten Mal statt
  - CF lehrt zum ersten Mal ein Bachelor-Pflichtfach
  - CF ist im Forschungssemester
  - Veranstaltung ist online
  - Prüfungsformat noch in der Schwebe
- Ein paar Bitten
  - Macht Euer Video an → auch wir Dozenten freuen uns über ein paar Gesichter
  - Stellt Fragen
  - Arbeitet mit, arbeitet nach

Unser Lehrprogramm im Bachelor



Agenda

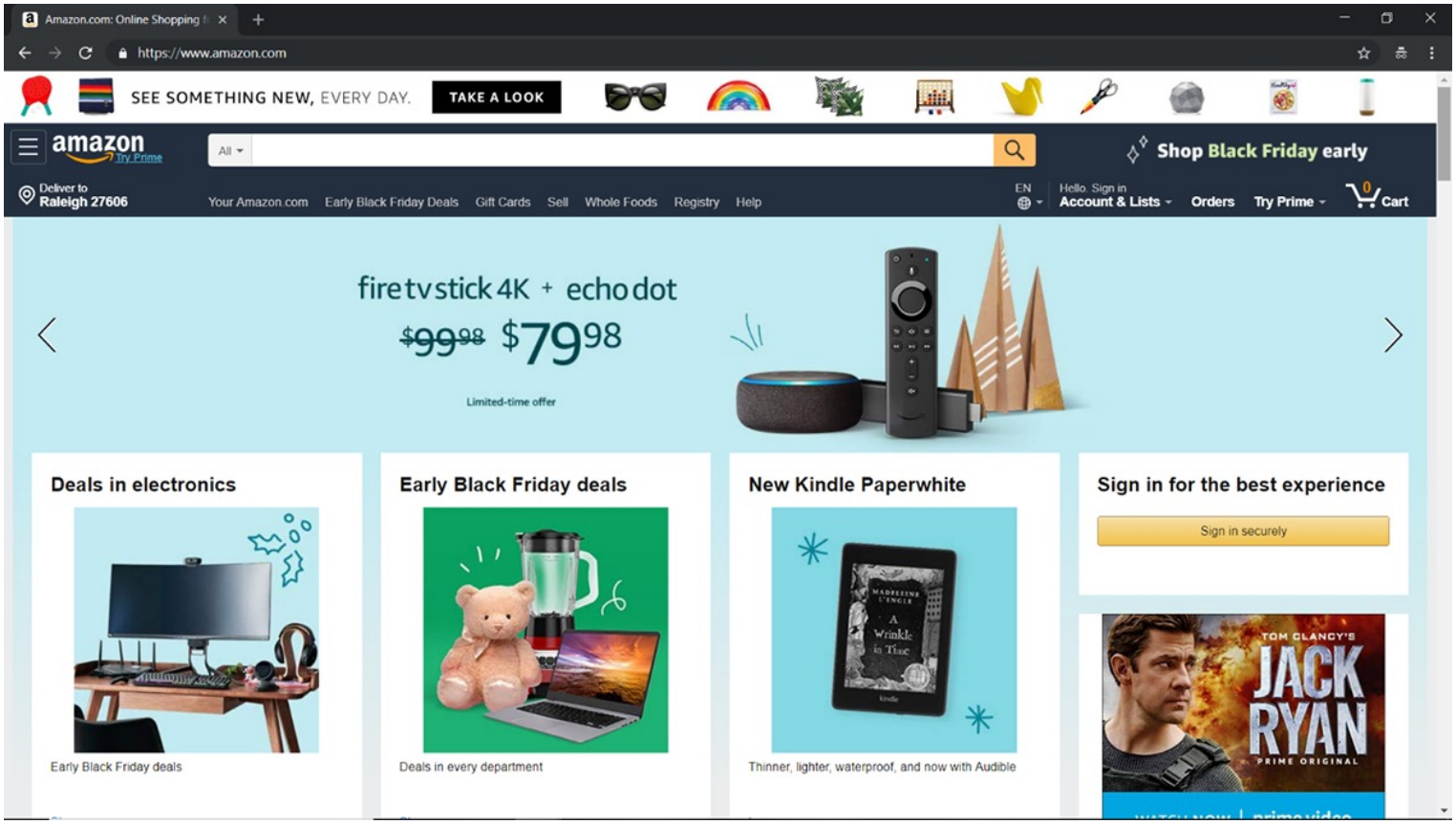
1. Organisatorisches
2. Motivation
3. Themenüberblick

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Überall Datenbanken, überall Daten

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

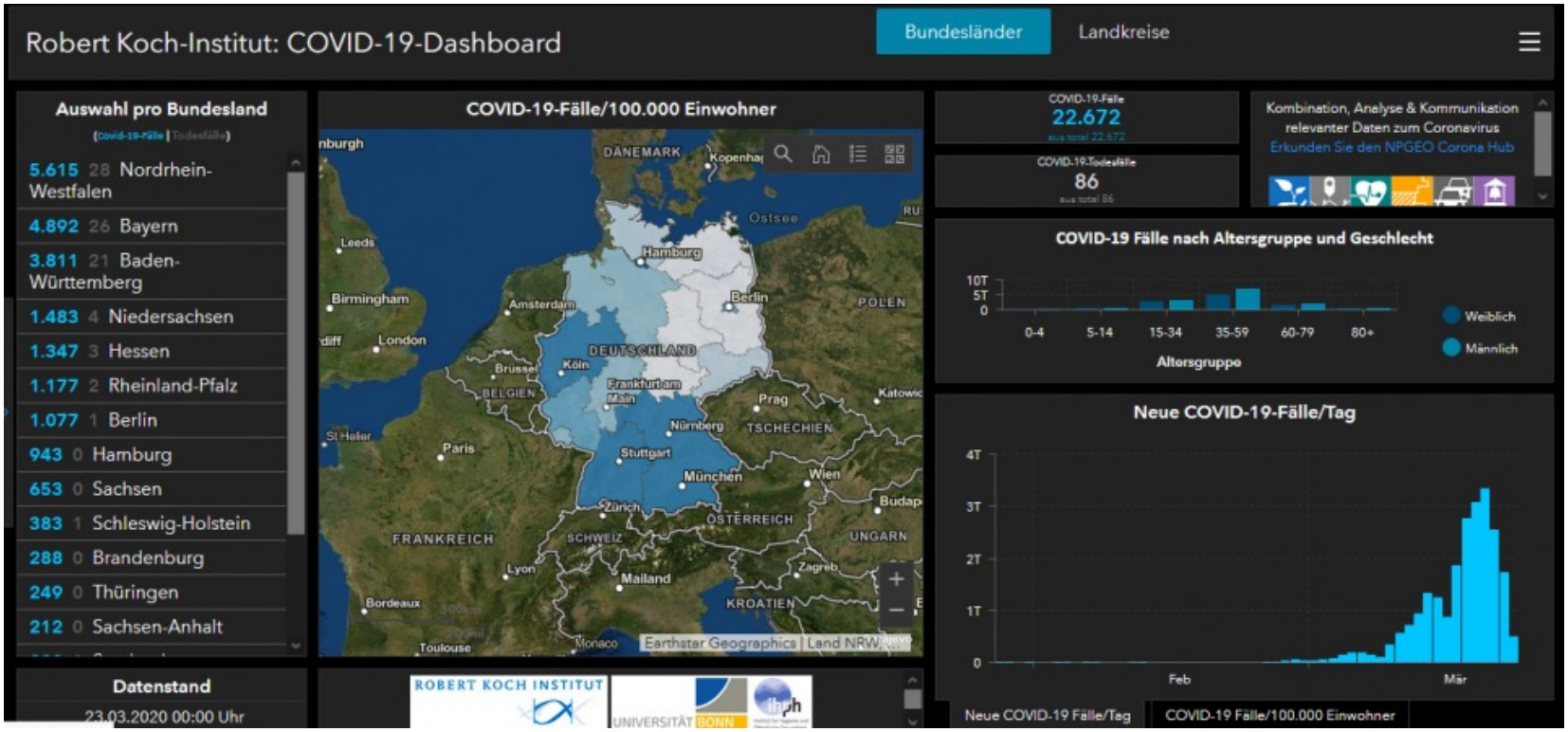
Überall Datenbanken, überall Daten



Überall Datenbanken, überall Daten

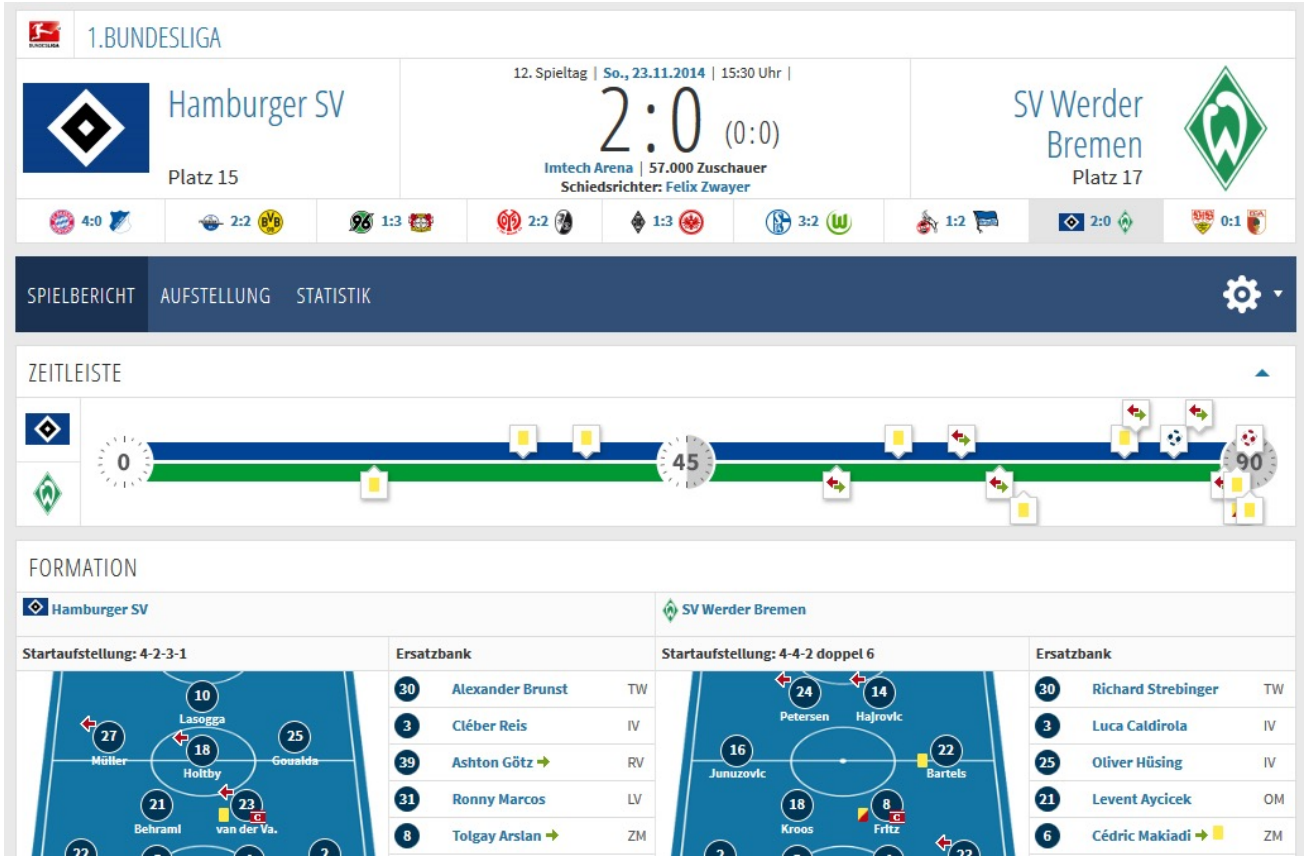
Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Überall Datenbanken, überall Daten



Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Überall Datenbanken, überall Daten



Sexiest Job of the 21st Century

- Daten sind der Rohstoff der 21. Jahrhunderts
- Daten stellen immensen Wert für jedes Unternehmen / jede Organisation dar
- Datenmanagement durchdringt jede Phase der Wertschöpfungskette und entscheidet letztendlich ob ein Unternehmen in der Zukunft als Gewinner weiter bestehen wird oder als Verlierer - Opfer der „Digital Disruption“ zugrunde geht!

Data Scientist

- **Job**
  Ein Data Scientist übernimmt im Grunde Aufgaben einer internen Unternehmensberatung. Er soll die gewaltigen Datenmengen in konkrete Handlungsanweisungen für das Unternehmen überführen. Dazu führt er Daten aus verschiedenen Abteilungen zusammen, baut daraus übergreifende Analysen und Modelle. Daten aufbereiten, Lösungen entwickeln und Ergebnisse präsentieren sind wichtige Aufgaben.
- **Voraussetzungen**
  Als Data Scientist haben Sie in der Regel ein Studium in Mathematik oder Informatik absolviert und am besten den Schwerpunkt Statistik gewählt. Auch Wirtschaftsmathematiker und Wirtschaftsingenieure sind Kandidaten. Mittlerweile gibt es an den Unis auch vermehrt Studiengänge in Data Science, in München und Darmstadt zum Beispiel. Statistiktools, Programmiersprachen und Datenbanksysteme sind wichtige Fähigkeiten.
- **Gehalt**
  Mit guter Ausbildung und im entsprechenden Unternehmen können zum Teil 60.000 Euro und mehr verdient werden – andere Stellen werden allerdings deutlich geringer entlohnt.

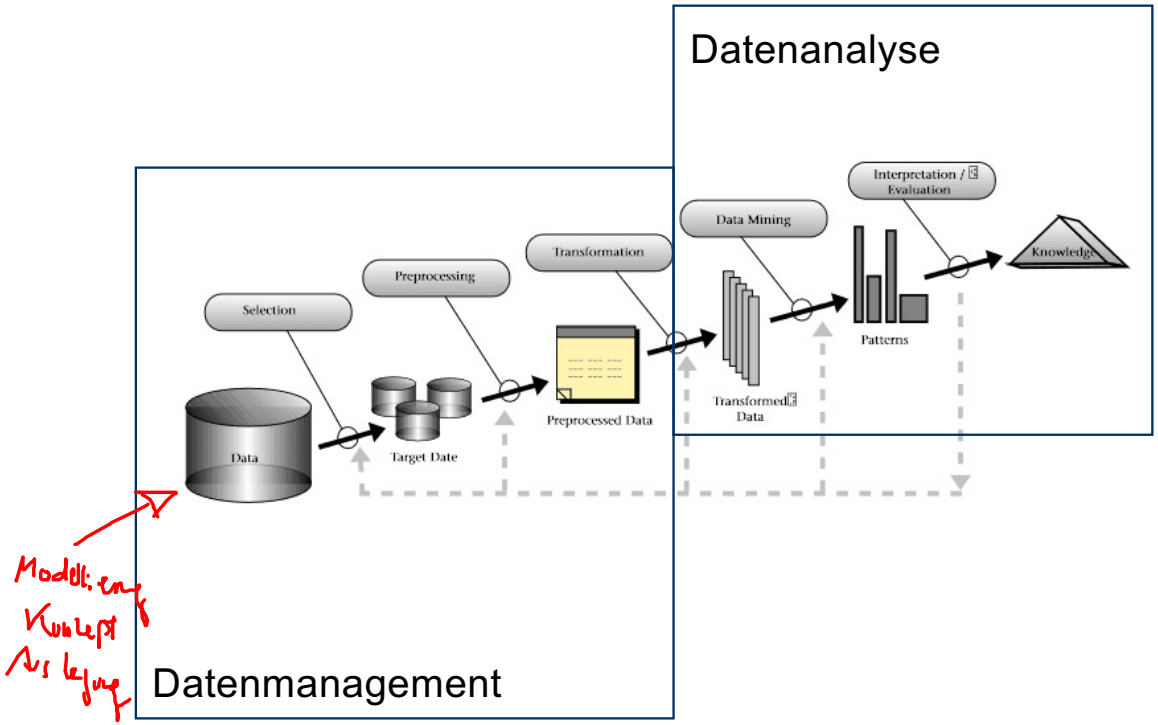
Irgendjemand muss den Job ja erledigen!

Agenda

1. Organisatorisches
2. Motivation
3. Themenüberblick

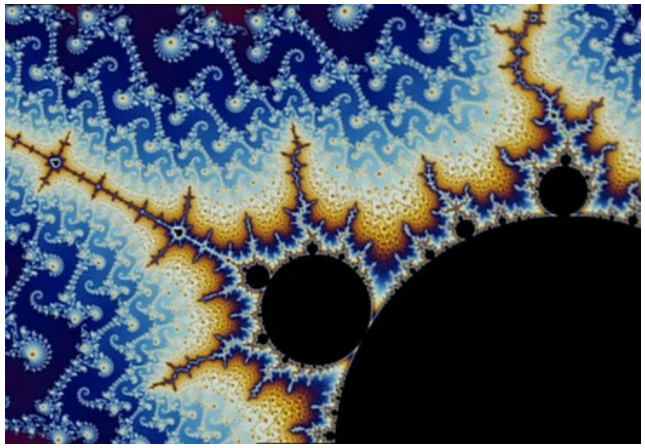
Vorläufiger Überblick der Themen

- Konzepte klassischer Datenbankarchitekturen
- Datenmodellierung und Normalisierung
- Einführung in eine Anfragesprache
- Nutzung von Datenbanken
- Hypothesengetriebene und modellbildende Datenanalyse
- Datenanalyseprozesse und deren Vergleich
- Überwachte und unüberwachte Lernverfahren
- Konzeption und Umsetzung (komplexer) Datenanalysen



Was sind eigentlich Daten?

```
Dateiname                       : IMG_4601.JPG
Kameramodell                    : Canon EOS 300D DIGITAL
Erstellungsdatum/-uhrzeit       : 2010:03:03 12:39:21
Aufnahmemodus                   : Programmautomatik
Belichtungsdauer                : 1/60
Blende                          : 4.0
Belichtungsmessmethode          : Mehrfeldmessung
Belichtungskorrektur            : 0
ISO-Empfindlichkeit             : 100
Objektiv                        : 18.0 - 55.0 mm
Brennweite                      : 22.0 mm
Bildgröße                       : 3072x2048
Bildqualität                    : Fein
Blitz                           : Blitz wurde ausgelöst,
                                  Blitz erzwingen-Modus,
                                  Rote-Augen-Reduzierung
Blitztyp                        : Intern
Blitzbelichtungsausgleich       : 0
Rote Augen Reduzierung          : Ein
Shutter Curtain Sync            : 1st-curtain sync
Weißabgleich                    : Automatisch
Fokus-Modus                     : Manueller Fokus (3)
Kontrast                        : Standard
Schärfe                         : 0
Farbsättigung                   : Standard
Farbton                         : Standard
Dateigröße                      : 1292 kB
Dateinummer                     : 146-4601
Aufnahmeart                     : Einzelbild
Name des Besitzers              :
Seriennummer                    : 1830536199
```



```
01010100 01101000 01101001 01110011
00100000 01101001 01110011 00100000
01110100 01101000 01100101 00100000
01110100 01110101 01110100 01101111
01110010 01101001 01100001 01101100
00100000 01110100 01101111 00100000
01101100 01100101 01100001 01110010
01101110 00100000 01100010 01101001
01101110 01100001 01110010 01111001
00101110 00100000 01001001 00100000
01101000 01101111 01110000 01100101
00100000 01111001 01101111 01110101
00100000 01100101 01101110 01101010
01101111 01111001 00100000 01101001
01110100 00100001
```

```
<?xml version="1.0"?>
<quiz>
 <qanda seq="1">
  <question>
   Who was the forty-second
   president of the U.S.A.?
  </question>
  <answer>
   William Jefferson Clinton
  </answer>
 </qanda>
 <!-- Note: We need to add
  more questions later.-->
</quiz>
```

XML

Datendefinitionen

- Unter Daten versteht man im Allgemeinen **Angaben**, (Zahlen-)Werte oder formulierbare Befunde, die durch Messung, Beobachtung u. a. gewonnen wurden

- Gebilde aus **Zeichen** oder kontinuierliche Funktionen, die aufgrund bekannter oder unterstellter Abmachungen **Informationen** darstellen, vorrangig zum **Zweck der Verarbeitung** und als deren **Ergebnis**

- Typische Unterscheidung

- Unstrukturiert
  - Dokumente, Grafiken, Töne

- Semistrukturiert
  - Extensible Markup Language (XML)

- **Strukturiert**
  - **Datenbanken**, Dateien

Wikipedia (2014), DIN (1985)

Es geht nicht ohne Datenbanken!

- Redundanz und Inkonsistenz
  - Vermeidung Mehrfachspeicherung, Aktualisierungsprobleme?
- Zugriffsbeschränkungen
  - Verknüpfung von Daten?
- Mehrbenutzerbetrieb
  - Lost Updates?
- Datenverlust
  - Backups?
- Integritätsverletzung
  - Abhängigkeiten?
- Sicherheitsprobleme
  - Zugriffsschutz?
- Entwicklungskosten
  - Keine Abstraktionsschicht

- (Unternehmens-)Daten gehören in Datenbanken
- Datenmanagement ist Datenbank-Management im erweiterten Sinne

Kemper, Eickler (2013)

Was verstehen wir unter Datenmanagement?

- Datenverwaltung – dehnbarer Begriff, aber versuchen wir es ...

- Das Ziel: ein real existierendes Problem so abzubilden, dass es mittels Datenerfassung und Datenauswertung simuliert, erklärt und verstanden werden kann.

- Relevante Teilaspekte:
  - Modellierung
  - Erfassung
  - Aufbewahrung
  - Aufbereitung und Zurverfügungstellung

Anforderungen an das Datenmanagement

Die Daten müssen
- „zur rechten Zeit“
- „am rechten“ Ort
- in geeigneter Form
- im notwendigen Umfang
- dem autorisierten Benutzer
- entsprechend den Sicherheitsanforderungen übermittelt werden.

Ausblick nächste Vorlesung

- Konzepte klassischer Datenbankarchitekturen
- Einführung Datenmodellierung
  - Architektur Integrierter Informationssysteme
  - ER Modellierung